|  |
| --- |
| \chapter{Results and Discussion} |
|  |

|  |
| --- |
| \label{ch:disc} |
|  |

|  |
| --- |
| % Será necessária Introdução |
|  |

|  |
| --- |
| % ao Capítulo e talvez divisão em tópicos? |
|  |

|  |
| --- |
| \subsection{Activity along time}\label{constDisc} |
|  |

|  |
| --- |
| Regular patterns of activity were observed along time |
|  |

|  |
| --- |
| in the scales of seconds, minutes, hours, days and months. |
|  |

|  |
| --- |
| Histograms in each of the time scales were computed as were circular average and dispersion values, and the results are given in Tables~\ref{tab:circ}-\ref{tab:min22}. For example, uniform activity is found with respect to seconds, minutes and days of the months. Weekend days exhibit about half the activity of regular weekdays, and there is a peak of activity between 11am and noon. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \begin{table} |
|  |

|  |
| --- |
| \caption{The rescaled circular mean $\theta\_\mu'$ and the circular dispersion $\delta(z)$, described in Section~\ref{sec:mtime}, for different timescales. This example table was constructed using all LAD messages, and the results are the same for other lists, as shown in Section~\ref\*{si:circ} of the Supporting Information document. The most uniform distribution of activity was found in seconds and minutes. Hours of the day exhibited the most concentrated activity (lowest $\delta(z)$), with mean between 2 p.m. and 3 p.m. ($\theta'=-9.61$). Weekdays, days of the month and months have mean near zero (i.e. near the beginning of the week, month and year) and high dispersion. Note that $\theta\_u'$ has the dimensional unit of the corresponding time period while $\delta(z)$ is dimensionless.} |
|  |

|  |
| --- |
| \begin{center} |
|  |

|  |
| --- |
| \begin{tabular}{ |l|| c|c| } |
|  |

|  |
| --- |
| \hline |
|  |

|  |
| --- |
| %scale & $\theta\_\mu'$ & $S(z)$ & $Var(z)$ & $\delta(z)$ & $\frac{max(incidence)}{min(incidence)}$ & $ \mu\_{\frac{max(incidence')}{min(incidence')}} $ & $ \sigma\_{\frac{max(incidence')}{min(incidence')} } $ \\ \hline\hline |
|  |

|  |
| --- |
| %& $\theta\_\mu'$ & $S(z)$ & $Var(z)$ & $\delta(z)$ \\ \hline\hline |
|  |

|  |
| --- |
| scale & mean $\theta\_\mu'$ & dispersion $\delta(z)$ \\ \hline |
|  |

|  |
| --- |
| \input{tables/tab2TimeLAD\_\_\_} |
|  |

|  |
| --- |
| \end{tabular} |
|  |

|  |
| --- |
| \begin{flushleft} |
|  |

|  |
| --- |
| Source: Prepared by the authors.\ |
|  |

|  |
| --- |
| \end{flushleft} |
|  |

|  |
| --- |
| \end{center} |
|  |

|  |
| --- |
| \label{tab:circ} |
|  |

|  |
| --- |
| \end{table} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \begin{table} |
|  |

|  |
| --- |
| \caption{Activity percentages along the hours of the day. Nearly identical distributions were observed on other social systems as shown in Section~\ref\*{si:hours} of the Supporting Information document. |
|  |

|  |
| --- |
| Highest activity was observed between noon and 6pm (with 1/3 of total day activity), followed by the time period between 6pm and midnight. |
|  |

|  |
| --- |
| Around 2/3 of the activity takes place from noon to midnight |
|  |

|  |
| --- |
| but the activity peak occurs between 11 a.m. and 12 p.m. |
|  |

|  |
| --- |
| This table shows results for the activity in CPP.} |
|  |

|  |
| --- |
| \footnotesize |
|  |

|  |
| --- |
| \input{tables/tabHoursCPP\_} |
|  |

|  |
| --- |
| \label{tab:hin} |
|  |

|  |
| --- |
| \begin{flushleft} |
|  |

|  |
| --- |
| Source: Prepared by the authors.\ |
|  |

|  |
| --- |
| \end{flushleft} |
|  |

|  |
| --- |
| \end{table} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \begin{table} |
|  |

|  |
| --- |
| \caption{Activity percentages along weekdays. |
|  |

|  |
| --- |
| Higher activity was observed during workweek days, with a decrease of activity on weekend days of at least one third and at most two thirds.} |
|  |

|  |
| --- |
| \begin{center} |
|  |

|  |
| --- |
| \begin{tabular}{ | l || c | c | c | c | c | c | c |} |
|  |

|  |
| --- |
| \hline |
|  |

|  |
| --- |
| & Mon & Tue & Wed & Thu & Fri & Sat & Sun \\ \hline |
|  |

|  |
| --- |
| \input{tables/tabWeekdays} |
|  |

|  |
| --- |
| \end{tabular} |
|  |

|  |
| --- |
| \end{center} |
|  |

|  |
| --- |
| \label{tab:win} |
|  |

|  |
| --- |
| \begin{flushleft} |
|  |

|  |
| --- |
| Source: Prepared by the authors.\ |
|  |

|  |
| --- |
| \end{flushleft} |
|  |

|  |
| --- |
| \end{table} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| In the scales of seconds and minutes, activity is uniform, |
|  |

|  |
| --- |
| with the messages being slightly more evenly distributed in all lists than in simulations with the uniform distribution\footnote{Numpy version 1.8.2, ``random.randint'' function, was used for simulations, algorithms in \url{https://github.com/ttm/percolation}.}. |
|  |

|  |
| --- |
| In the networks, $\frac{min(incidence)}{max(incidence)} \in (0.784,.794)$ while simulations reach these values but have on average more discrepant higher and lower peaks, i.e. if $\xi=\frac{min(incidence')}{max(incidence')}$ than $\mu\_\xi=0.7741 \text{ and } \sigma\_\xi=0.02619$. |
|  |

|  |
| --- |
| Therefore, the incidence of messages at each second of a minute and at each minute of an hour was considered uniform. |
|  |

|  |
| --- |
| In these cases, the circular dispersion is maximized and the mean has little meaning as indicated in Table~\ref{tab:circ}. |
|  |

|  |
| --- |
| As for the hours of the day, an abrupt peak is found between 11am and 12pm with the most active period being the afternoon, with one third of total daily activity, and two thirds of activity are allocated in the second 12h of each day. Days of the week revealed a decrease between one third and two thirds of activity on weekends. |
|  |

|  |
| --- |
| Days of the month were regarded as homogeneous with an inconclusive slight tendency of the first week to be more active. |
|  |

|  |
| --- |
| Months of the year revealed patterns matching usual work and academic calendars. The time period examined here was not sufficient for the analysis of activity along the years. These patterns are exemplified in Tables~\ref{tab:hin}-\ref{tab:min22}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \FloatBarrier |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \begin{table} |
|  |

|  |
| --- |
| \caption{Activity along the days of the month cycle. |
|  |

|  |
| --- |
| Nearly identical distributions are found in all systems |
|  |

|  |
| --- |
| as indicated in Section~\ref\*{si:monthdays} of the Supporting Information. Although slightly higher activity rates are found in the beginning of the month, the most important feature seems to be the homogeneity made explicit by the high circular dispersion in Table~\ref{tab:circ}. |
|  |

|  |
| --- |
| This specific example and empirical table correspond to the activity of the MET email list.} |
|  |

|  |
| --- |
| \footnotesize |
|  |

|  |
| --- |
| \input{tables/tabMonthdaysMET} |
|  |

|  |
| --- |
| \label{tab:min} |
|  |

|  |
| --- |
| \begin{flushleft} |
|  |

|  |
| --- |
| Source: Prepared by the authors.\ |
|  |

|  |
| --- |
| \end{flushleft} |
|  |

|  |
| --- |
| \end{table} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \begin{table} |
|  |

|  |
| --- |
| \caption{Activity percentages on months along the year. Activity is usually concentrated in Jun-Aug and/or in Dec-Mar, potentially due to academic calendars, vacations and end-of-year holidays. This table corresponds to activity in LAU. Similar results are shown for other lists in Section~\ref\*{si:months} of the Supporting Information document.} |
|  |

|  |
| --- |
| \footnotesize |
|  |

|  |
| --- |
| \input{tables/tabMonthsLAU} |
|  |

|  |
| --- |
| \label{tab:min22} |
|  |

|  |
| --- |
| \begin{flushleft} |
|  |

|  |
| --- |
| Source: Prepared by the authors.\ |
|  |

|  |
| --- |
| \end{flushleft} |
|  |

|  |
| --- |
| \end{table} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Stable sizes of Erd\"os sectors}\label{subsec:pih} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| The distribution of vertices in the hub, intermediary and periphery Erd\"os sectors is remarkably stable along time if the snapshots hold 200 or more messages, |
|  |

|  |
| --- |
| as it is clear in Figure~\ref{fig:sectIL} and in Section~\ref{si:frac} of the Appendix. |
|  |

|  |
| --- |
| %Moreover, all email lists analyzed exhibit the same distribution profile. |
|  |

|  |
| --- |
| Activity is highly concentrated on the hubs, while a very large number of peripheral vertices contribute to only a fraction of the activity. |
|  |

|  |
| --- |
| This is expected for a system with a scale-free profile, as confirmed with the distribution of activity among participants in Table~\ref{autores}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Typically, $[3\%-12\%]$ of the vertices are hubs, |
|  |

|  |
| --- |
| $[15\%-45\%]$ are intermediary and $[44\%-81\%]$ are peripheral, |
|  |

|  |
| --- |
| which is consistent with other studies~\cite{secFree}. |
|  |

|  |
| --- |
| These results hold for the total, in and out degrees and strengths. |
|  |

|  |
| --- |
| Stable sizes are also observed for 100 or less messages if the classification |
|  |

|  |
| --- |
| of the three sectors is performed with one of the compound criteria established in Section~\ref{sectioning}. The networks often hold this basic structure with as few as 10-50 messages, i.e. concentration of activity and the abundance of low-activity participants take place even with very few messages, which is highlighted in Section~\ref{si:frac} of the Appendix. |
|  |

|  |
| --- |
| A minimum window size for the observation of more general properties might be inferred by monitoring |
|  |

|  |
| --- |
| both the giant component and the degeneration of the Erd\"os sectors. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| In order to support the generality of these findings, |
|  |

|  |
| --- |
| we list the Erd\"os sector sizes of 12 networks from Facebook, Twitter and Participabr in Table~\ref{tab:secE} of the Appendix. |
|  |

|  |
| --- |
| The fractions of hubs, intermediary and peripheral nodes are |
|  |

|  |
| --- |
| essentially the same as for the email list networks but with exceptions and a greater variability. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \begin{figure\*} |
|  |

|  |
| --- |
| \centering |
|  |

|  |
| --- |
| \caption{Stability of Erd\"os sector sizes. |
|  |

|  |
| --- |
| Fractions of participants derived from degree and strength criteria, $E\_1$ and $E\_4$ described in Section~\ref{sectioning}, are both on the left. |
|  |

|  |
| --- |
| Fractions derived from the exclusivist $C\_1$ and the inclusivist $C\_2$ compound criteria are shown in the plots to the right. |
|  |

|  |
| --- |
| The ordinates $\overline{e\_{\gamma,\phi}}=\frac{|e\_{\gamma,\phi}|}{N}$ denote the fraction of participants in sector $\phi$ through criterion $E\_\gamma$ |
|  |

|  |
| --- |
| and, similarly, $\overline{c\_{\delta,\phi}}=\frac{|c\_{\delta,\phi}|}{N}$ denotes the fraction of participants in sector $\phi$ through criterion $C\_\delta$. |
|  |

|  |
| --- |
| Sections~\ref\*{si:frac} and~\ref\*{si:ext} of the Supporting Information bring a systematic collection of such timeline figures with all simple and compound criteria specified in Section~\ref{sectioning}, with results for networks from Facebook, Twitter and Participabr.} |
|  |

|  |
| --- |
| \includegraphics[width=\textwidth]{figs/InText-WLAU-S1000\_\_} |
|  |

|  |
| --- |
| \begin{flushleft} |
|  |

|  |
| --- |
| Source: Prepared by the authors.\ |
|  |

|  |
| --- |
| \end{flushleft} |
|  |

|  |
| --- |
| \label{fig:sectIL} |
|  |

|  |
| --- |
| \end{figure\*} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \begin{table}[h] |
|  |

|  |
| --- |
| \caption{Distribution of activity among participants. |
|  |

|  |
| --- |
| The first column shows the percentage of messages sent by the most active participant. The column for the first quartile ($Q\_1$) gives the minimum percentage of participants responsible for at least 25\% of total messages with the actual percentage in parentheses. Similarly, the column for the first three quartiles $Q\_3$ gives the minimum percentage of participants responsible for 75\% of total messages. |
|  |

|  |
| --- |
| The last decile $D\_{-1}$ column shows the maximum percentage of participants responsible for 10\% of messages.} |
|  |

|  |
| --- |
| \begin{center} |
|  |

|  |
| --- |
| \begin{tabular}{ | l || c | c | c | c | } |
|  |

|  |
| --- |
| \hline |
|  |

|  |
| --- |
| list & hub & $ Q\_1 $ & $ Q\_3 $ & $D\_{-1}$ \\ \hline |
|  |

|  |
| --- |
| \input{tables/userTab} |
|  |

|  |
| --- |
| \end{tabular} |
|  |

|  |
| --- |
| \end{center} |
|  |

|  |
| --- |
| \begin{flushleft} |
|  |

|  |
| --- |
| Source: Prepared by the authors.\ |
|  |

|  |
| --- |
| \end{flushleft} |
|  |

|  |
| --- |
| \label{autores} |
|  |

|  |
| --- |
| \end{table} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Stability of principal components}\label{prevalence} |
|  |

|  |
| --- |
| %The topology was analyzed using standard, well-established metrics of centrality and clustering. |
|  |

|  |
| --- |
| %We also introduced symmetry metrics given the evidence of their importance in social contexts~\cite{newmanEvolving}. |
|  |

|  |
| --- |
| %The contribution of each metric to the variance is very similar for all the networks and along time. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| The principal components of the participants are very stable in the topological space, |
|  |

|  |
| --- |
| i.e. in the space of network metrics. |
|  |

|  |
| --- |
| Table~\ref{tab:pcainin} exemplifies the formation of principal components by providing the averages over non-overlapped activity snapshots of a network. The most important result of this application of PCA, the stability of principal components, is underpinned by the very small dispersion of the contribution of each metric to each principal component. |
|  |

|  |
| --- |
| %The contribution of each metric to the |
|  |

|  |
| --- |
| %principal components presents |
|  |

|  |
| --- |
| %very small standard deviation. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \begin{table}[!h] |
|  |

|  |
| --- |
| \caption{Loadings for the 14 metrics into the principal components for the MET list, $1000$ messages in 20 disjoint positions. The clustering coefficient (cc) appears as the first metric in the table, followed by 7 centrality metrics and 6 symmetry-related metrics. Note that the centrality measurements, including degrees, strength and betweenness centrality, are the most important contributors for the first principal component, while the second component is dominated by symmetry metrics. The clustering coefficient is only relevant for the third principal component. The three components have in average more than 85\% of the variance. |
|  |

|  |
| --- |
| The low standard deviation $\sigma$ implies that the principal components are considerably stable.} |
|  |

|  |
| --- |
| \footnotesize |
|  |

|  |
| --- |
| \input{tables/tabPCA3CPP} |
|  |

|  |
| --- |
| \label{tab:pcainin} |
|  |

|  |
| --- |
| \begin{flushleft} |
|  |

|  |
| --- |
| Source: Prepared by the authors.\ |
|  |

|  |
| --- |
| \end{flushleft} |
|  |

|  |
| --- |
| \end{table} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| The first principal component is an average of centrality metrics: |
|  |

|  |
| --- |
| degrees, strengths and betweenness centrality. |
|  |

|  |
| --- |
| On one hand, the similar relevance of all centrality metrics is not surprising since they are highly correlated, |
|  |

|  |
| --- |
| e.g. degree and strength have Spearman correlation coefficient $\in [0.95,1]$ |
|  |

|  |
| --- |
| and Pearson coefficient $\in [0.85,1)$ for window sizes greater than a thousand messages. |
|  |

|  |
| --- |
| On the other hand, each of these metrics is related to a different participation characteristic, |
|  |

|  |
| --- |
| and their equal relevance for variability, |
|  |

|  |
| --- |
| as measured by the principal component, is noticeable. |
|  |

|  |
| --- |
| Also, this suggests that these centrality metrics |
|  |

|  |
| --- |
| are equally adequate for characterizing the networks |
|  |

|  |
| --- |
| and the participants. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \begin{figure} |
|  |

|  |
| --- |
| \centering |
|  |

|  |
| --- |
| %\includegraphics[width=.6\textwidth,height=10cm]{figs/im13PCAPLOT\_\_} |
|  |

|  |
| --- |
| \caption{The first plot highlights the well-known pattern of degree versus clustering coefficient, characterized by the higher clustering coefficient of lower degree vertices. |
|  |

|  |
| --- |
| The second plot shows the greater dispersion of the symmetry-related ordinates dominant in the second principal component (PC2). |
|  |

|  |
| --- |
| This larger dispersion suggests that symmetry-related metrics are more powerful, |
|  |

|  |
| --- |
| for characterizing interaction networks than the clustering coefficient, |
|  |

|  |
| --- |
| especially for hubs and intermediary vertices. |
|  |

|  |
| --- |
| This figure reflects a snapshot of the LAU list with 1000 contiguous messages.} |
|  |

|  |
| --- |
| \includegraphics[width=.45\textwidth]{figs/im13PCAPLOT\_\_} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| % Similar structures were observed in all window sizes $ws\;\in\;[500,10000]$, in networks derived from email lists, |
|  |

|  |
| --- |
| % and in networks from Facebook, Twitter and Participabr, |
|  |

|  |
| --- |
| % which suggests a common relationship between the metrics of degrees, strengths and betweenness centrality, |
|  |

|  |
| --- |
| % the symmetry-related metrics and clustering coefficient.} |
|  |

|  |
| --- |
| \label{fig:sym} |
|  |

|  |
| --- |
| \begin{flushleft} |
|  |

|  |
| --- |
| Source: Prepared by the authors.\ |
|  |

|  |
| --- |
| \end{flushleft} |
|  |

|  |
| --- |
| \end{figure} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| According to Table~\ref{tab:pcainin} and Figure~\ref{fig:sym}, |
|  |

|  |
| --- |
| dispersion is larger in symmetry-related metrics than in clustering coefficient. |
|  |

|  |
| --- |
| % As expected from basic complex network theory, peripheral vertices have low values of centrality metrics and larger dispersion with regard to the clustering coefficient. |
|  |

|  |
| --- |
| % %The scatter plot in the third system of Figure~\ref{fig:sym}, |
|  |

|  |
| --- |
| % %where all metrics are considered and there is a greater dispersion |
|  |

|  |
| --- |
| % %with respect to the ordinates, |
|  |

|  |
| --- |
| % This reflects in the relevance of the symmetry-related metrics. |
|  |

|  |
| --- |
| We conclude that the symmetry metrics are more powerful, in terms of dispersion in the topological metrics space, in characterizing interaction networks and their participants, than the clustering coefficient, especially for hubs and intermediary vertices (peripheral vertices have larger dispersion with regard to the clustering coefficient). |
|  |

|  |
| --- |
| Interestingly, the clustering coefficient is always combined |
|  |

|  |
| --- |
| with the standard deviation of the asymmetry and disequilibrium |
|  |

|  |
| --- |
| of edges $\sigma^{asy}$ and $\sigma^{dis}$ in the third principal component. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| %These results are also reported for 12 networks from Facebook, Twitter and Participabr |
|  |

|  |
| --- |
| %in Section~\ref{si:ext} of the Supporting Information document. |
|  |

|  |
| --- |
| Similar results are presented in Sections~\ref{si:pcat} and~\ref{si:ext} |
|  |

|  |
| --- |
| of the Appendix for other email lists and interaction networks. A larger variability was found for the latter networks, |
|  |

|  |
| --- |
| which motivated the use of interaction networks derived from email lists for benchmarking. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| %the overall behavior was maintained in that centrality measurements |
|  |

|  |
| --- |
| %were found prevalent in the first principal component, |
|  |

|  |
| --- |
| %followed by symmetry-related metrics on the second principal |
|  |

|  |
| --- |
| %component and then clustering coefficient on the third principal component. |
|  |

|  |
| --- |
| %Similar results are presented in Sections~\ref{si:pcat} and~\ref{si:ext} |
|  |

|  |
| --- |
| %of the Supporting Information document for other email lists and other interaction networks, |
|  |

|  |
| --- |
| %with the consideration of strategic combinations of metrics. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Types from Erd\"os sectors}\label{sec:pty} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Assigning a type to a participant raises important issues about the scientific cannon for human types and the potential for stigmatization and prejudice. The Erd\"os sector to which a participant belongs can be regarded as implying a social type for this participant. |
|  |

|  |
| --- |
| In this case, the type of a participant changes both along time and as different networks are considered, despite the stability of the network. Therefore, the potential for prejudice of such participant typology is attenuated~\cite{adorno}. In other words, an individual is a hub in a number of networks and peripheral in other networks, and even within the same network he/she most probably changes type along time~\cite{animacoes}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| The importance of this issue can be grasped by the consideration of static types derived from quantitative criteria. For example, in email lists with a small number of participants, the number of threads has a negative correlation with the number of participants. |
|  |

|  |
| --- |
| When the number of participants exceeds a threshold, the number of threads has a positive correlation with the number of participants. |
|  |

|  |
| --- |
| This finding is illustrated in Figure~\ref{fig:nmgamma3d} |
|  |

|  |
| --- |
| and can also be observed in Table~\ref{tab:genLists}. |
|  |

|  |
| --- |
| The assignment of types to individuals, in this latter case, |
|  |

|  |
| --- |
| has more potential for prejudice because |
|  |

|  |
| --- |
| the derived participant type is static and |
|  |

|  |
| --- |
| one fails to acknowledge that |
|  |

|  |
| --- |
| human individuals are not immutable entities. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Further observations regarding the Erd\"os sectors |
|  |

|  |
| --- |
| and the implicit participant types were made, which are consistent with the literature~\cite{barabasiEvo}: 1) hubs and intermediary participants usually have intermittent activity, and stable activity was found only in smaller communities. For instance, the MET list had stable hubs while LAU, LAD and CPP exhibited intermittent hubs. |
|  |

|  |
| --- |
| 2) Network structure seems to be most influenced by the |
|  |

|  |
| --- |
| activity of intermediary participants as they have less extreme |
|  |

|  |
| --- |
| roles than hubs and peripheral participants and |
|  |

|  |
| --- |
| can therefore connect to the sectors and other participants |
|  |

|  |
| --- |
| in a more selective and explicit manner. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| %Moreover, such typology of participants bridges exact and human sciences and may |
|  |

|  |
| --- |
| %be enriched with concepts from other typologies, |
|  |

|  |
| --- |
| %such as Meyer-Briggs, Pavlov or the authoritarian types of the F-Scale~\cite{adorno}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| %We analyzed the temporal evolution of the networks |
|  |

|  |
| --- |
| %using visualization |
|  |

|  |
| --- |
| %tools developed for this research~\cite{rcText,versinus} |
|  |

|  |
| --- |
| %and inspected raw data. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| %dictated (or revealed) by the |
|  |

|  |
| --- |
| %(e.g. stable or intermittent patterns of activity and preferential communication |
|  |

|  |
| --- |
| %with hubs or periphery) |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| %of both hubs and peripheral vertices |
|  |

|  |
| --- |
| %have the trivial facets of interacting |
|  |

|  |
| --- |
| % |
|  |

|  |
| --- |
| % |
|  |

|  |
| --- |
| % |
|  |

|  |
| --- |
| %\begin{itemize} |
|  |

|  |
| --- |
| % \item Typically, the activity of hubs is trivial: they interact as much as possible, in every occasion with everyone. |
|  |

|  |
| --- |
| %The activity of peripheral vertices also follows a simple pattern: they interact very rarely, in very few occasions. |
|  |

|  |
| --- |
| %Therefore, intermediary vertices seem responsible for the network structure. |
|  |

|  |
| --- |
| %Intermediary vertices may exhibit preferential communication to peripheral, intermediary, or hub vertices; can be marked by stable communication partners; can involve stable or intermittent patterns of activity, to point just a few examples of this greater variety of roles. |
|  |

|  |
| --- |
| %% \item Some of the most active participants receive many responses with relative few messages sent, and rarely are top hubs. |
|  |

|  |
| --- |
| %%These seem as authorities and contrast with participants that respond much more than receive responses. |
|  |

|  |
| --- |
| %% \item The most obvious community structure, as observed by a high clustering coefficient, i.e. members know each other often, is found mostly in peripheral and intermediary sectors. |
|  |

|  |
| --- |
| %\end{itemize} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| %Within networks as the whole objects of analysis, |
|  |

|  |
| --- |
| %we were able to observe a peculiar correlation pattern |
|  |

|  |
| --- |
| %between the number of threads and the number of participants. |
|  |

|  |
| --- |
| \begin{figure} |
|  |

|  |
| --- |
| \centering |
|  |

|  |
| --- |
| \caption{A scatter plot of number of messages $M$ versus number of participants $N$ versus number of threads $\Gamma$ for 140 email lists. |
|  |

|  |
| --- |
| Highest $\Gamma$ is associated with low $N$. |
|  |

|  |
| --- |
| The correlation between $N$ and $\Gamma$ is negative for low values of $N$ but positive otherwise. |
|  |

|  |
| --- |
| This negative correlation between $N$ and $\Gamma$ can also be observed in Table~\ref{tab:genLists}. |
|  |

|  |
| --- |
| Accordingly, for $M=20000$ messages, this inflection |
|  |

|  |
| --- |
| of correlation was found around $N=1500$, while CPP, LAU, LAD, MET lists |
|  |

|  |
| --- |
| present smaller networks.} |
|  |

|  |
| --- |
| \includegraphics[trim={0 0 0 1cm},clip,width=.7\columnwidth]{figs/mpgamma2\_} |
|  |

|  |
| --- |
| \begin{flushleft} |
|  |

|  |
| --- |
| Source: Prepared by the authors.\ |
|  |

|  |
| --- |
| \end{flushleft} |
|  |

|  |
| --- |
| \label{fig:nmgamma3d} |
|  |

|  |
| --- |
| \end{figure} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| %\section{Discussion} |
|  |

|  |
| --- |
| % given the results, and before reaching the conclusions |
|  |

|  |
| --- |
| % what to say? |
|  |

|  |
| --- |
| % --> what is the overall knowledge derived from the results |
|  |

|  |
| --- |
| % --> what are the limitations of this knowledge and of individual results |
|  |

|  |
| --- |
| % --> how should this results carry on is on the next sections. |
|  |

|  |
| --- |
| %\subsection{Consecutive scientific research} |
|  |

|  |
| --- |
| % --> research |
|  |

|  |
| --- |
| % textual diferences |
|  |

|  |
| --- |
| % audiovisualization of data |
|  |

|  |
| --- |
| % typologies, sociological critical theory, social psychology |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| %\subsection{Technological applications} |
|  |

|  |
| --- |
| % --> technological |
|  |

|  |
| --- |
| % resources categorization and recommendation |
|  |

|  |
| --- |
| % document creation |
|  |

|  |
| --- |
| % ontologies for the semantic web |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| %\subsection{Experimental and theoretical aspects of the research} |
|  |

|  |
| --- |
| % --> methods |
|  |

|  |
| --- |
| % Exploratory? |
|  |

|  |
| --- |
| % Hypothesis testing? |
|  |

|  |
| --- |
| % --> contributions |
|  |

|  |
| --- |
| % verifiable |
|  |

|  |
| --- |
| % knowledge |
|  |

|  |
| --- |
| % contextualization in the academic knowledge |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Implications of the main findings}\label{sec:impl} |
|  |

|  |
| --- |
| The findings reported in this thesis arose from an exploratory procedure to visually inspect the networks and to analyze considerable amounts of interaction networks data. |
|  |

|  |
| --- |
| % deriving from email lists and also from other networks. |
|  |

|  |
| --- |
| While this procedure has certainly an ad hoc nature, the statistics in the data are sufficiently robust for important features from these interaction networks to be extracted. |
|  |

|  |
| --- |
| Temporal stability, in the sense that interaction networks could be considered as stationary time series, is the most important feature. Also relevant is the significant stability found on the principal components, on the fraction of participants in each Erd\"os Sector and on the activity along different timescales. In fact, these findings confirm our initial hypothesis - based on the literature~\cite{newmanBook} - that interaction networks should exhibit some stability traces. The potential generality of these findings is suggested by the analysis of networks derived from diverse systems, with interaction networks from public email lists serving as proper benchmarks. Indeed, with such benchmarks one can compare any social network system. Furthermore, this analysis enables us to establish an outline of human interaction networks. It takes the hub, intermediary and periphery sectors out of the scientific folklore and into classes drawn from quantitative criteria. It enables the conception of non-static human types derived from natural properties. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| We envisage that the knowledge generated in the analysis may be exploited in applications where the type of each participant and the relative proportion of participants in each sector can be useful metadata. Just by way of illustration, this could be applied in semantic web initiatives, given that the Erd\"os sectorialization is static in a given snapshot. These results are also useful for classifying resources, e.g. in social media, and for resources recommendation to users~\cite{opa}. |
|  |

|  |
| --- |
| Finally, the knowledge acquired with a quantitative treatment of the whole data may help guide the creation through collective processes of documents to assist in participatory democracy. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Perhaps the most outreaching implications are related to sociological consequences. The results expose a classification of human individuals which is directly related to the concentration of wealth and based on natural laws. The derived human typology changes over different systems and over time in the same system, which implies a negation of the absolute concentration of wealth. Such concentration exists but changes across different wealth criteria and with time. Also, the hubs stand out as dedicated, sometimes enslaved, |
|  |

|  |
| --- |
| components of the social system. The peripheral participants have very limited interaction with the network. This suggests that intermediary participants tend to dictate structure, legitimate the hubs and stand out as authorities. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| With regard to the limitations of our study, one should emphasize that not all types of human interaction networks were analyzed. Therefore, the plausible generalization of properties has to be treated with caution, as a natural tendency of such systems and not as a rule. Also, the stable properties in the networks were not explored to the limit, which leaves many open questions. For example, what are the maximum and minimum sizes of the networks for which they hold? What is the outcome of PCA when more metrics are considered? What is the granularity in which the activity along the timescales is preserved? Do the findings reported also apply to other systems, beyond human networks? |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| % TTM find |
|  |

|  |
| --- |
| \section{Text-related results and discussion}\label{sec:tresults} |
|  |

|  |
| --- |
| % Aqui será necessária uma Introdução a este novo tipo de resultado. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| The most important result of including textual metrics in our analysis is the |
|  |

|  |
| --- |
| statistical evidence of extreme differentiation of each Erd\"os sector with respect to the texts produced. |
|  |

|  |
| --- |
| This conclusion can be reached by observing the differences in the measurements of textual features in |
|  |

|  |
| --- |
| each sector, but is reached with greater theoretical background from the adaptation of the |
|  |

|  |
| --- |
| Kolmogorov-Smirnov test we presented in Section~\ref{sec:ks}. |
|  |

|  |
| --- |
| % derived from the kolmogorov-smirnov test adaptation |
|  |

|  |
| --- |
| Other relevant results are: |
|  |

|  |
| --- |
| \begin{itemize} |
|  |

|  |
| --- |
| \item the achievement of references for the amount of nouns, adverbs, sizes of words, depth of Wordnet synsets and other linguistic traces, used in social networks. |
|  |

|  |
| --- |
| We did not find in literature any indication for such values and understood useful to acknowledge e.g. that about 15\% of the characters are spaces and more than 50\% of tokens are nouns. |
|  |

|  |
| --- |
| These values are available in~\cite{textTables} and not in the body of the text, |
|  |

|  |
| --- |
| since we focus in this thesis in the evidence that the texts from distinct sectors differ. |
|  |

|  |
| --- |
| \item Indicatives of what is different in the texts produced by each of the Erd\"os sectors. |
|  |

|  |
| --- |
| For example: hubs were found to use more contractions, |
|  |

|  |
| --- |
| more common words, and less punctuation if compared to the rest of the network, |
|  |

|  |
| --- |
| especially the peripheral sector. |
|  |

|  |
| --- |
| In general, the rise or fall of a text-related metric is not relevant or is monotonic along connectivity, |
|  |

|  |
| --- |
| but some of them reaches extreme values in the intermediary sector. |
|  |

|  |
| --- |
| % Comparative insights by means of statistics from email lists, the King James Bible and Shakespeare, which are often regarded as pillars of the English language. TTM |
|  |

|  |
| --- |
| \end{itemize} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| The next sections summarize results of immediate interest |
|  |

|  |
| --- |
| and further insights can be obtained by skimming through |
|  |

|  |
| --- |
| the tables and figures of~\cite{stab,textTables} and the Conclusions chapter. |
|  |

|  |
| --- |
| We illustrate with just one table of each kind, |
|  |

|  |
| --- |
| and from networks obtained with 2000 messages. |
|  |

|  |
| --- |
| In~\cite{textTables} we display tables for various networks. |
|  |

|  |
| --- |
| The scale of 1000 and 2000 messages was chosen for deriving results, |
|  |

|  |
| --- |
| as the networks in this scale are found with stable topological structure, |
|  |

|  |
| --- |
| as exposed in Section~\ref{subsec:pih}. |
|  |

|  |
| --- |
| The findings with 1000 messages are the same as with 2000 messages, |
|  |

|  |
| --- |
| and there are many tables. |
|  |

|  |
| --- |
| This motivated the exclusion of the tables with measurements in |
|  |

|  |
| --- |
| networks of 1000 messages. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{General characteristics of activity distribution among sectors}\label{sec:gen} |
|  |

|  |
| --- |
| Before we dig into the findings derived from text-related measures, |
|  |

|  |
| --- |
| let us look once more at the general structure of these networks, |
|  |

|  |
| --- |
| now giving emphasis on the activity of the sectors. |
|  |

|  |
| --- |
| In almost all our observations, |
|  |

|  |
| --- |
| the peripheral sector is responsible for starting most of the discussion threads, |
|  |

|  |
| --- |
| i.e. messages to the list which are not replies. |
|  |

|  |
| --- |
| This is surprising since the peripheral sector is responsible for fewer messages. |
|  |

|  |
| --- |
| This suggests a complementarity between peripheral diversity and hub specialization |
|  |

|  |
| --- |
| which, on its turn, deepens the understanding of the interaction network as a meaningful system. |
|  |

|  |
| --- |
| These assertions are condensed in Table~\ref{geralListas}. |
|  |

|  |
| --- |
| Less often, the intermediary sector is responsible for the greatest number of messages |
|  |

|  |
| --- |
| and of threads. |
|  |

|  |
| --- |
| Also meaningful is that the hubs sector is responsible for most of the messages, |
|  |

|  |
| --- |
| which is not completely obvious: hub participants are far more active but way less numbered. |
|  |

|  |
| --- |
| Interestingly, in such a setting where every characteristic differs with respect to |
|  |

|  |
| --- |
| distinct sectors, there was no evidence of difference on the size of the threads started by each sector. |
|  |

|  |
| --- |
| \input{tables/geralInline\_} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| %Also, comparing lists with a fixed number of messages, |
|  |

|  |
| --- |
| %the number of threads created seem to increase as the number of participants decrease. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Evidence that the texts from Erd\"os sectors differ}\label{subsec:di} |
|  |

|  |
| --- |
| Results from our adaptation of the Kolmogorov-Smirnov test (see Section~\ref{sec:ks}) |
|  |

|  |
| --- |
| present substantial evidence |
|  |

|  |
| --- |
| that the texts produced by each sector are different. |
|  |

|  |
| --- |
| Tables~\ref{tab:kolTok}-\ref{tab:kolPctInter} |
|  |

|  |
| --- |
| illustrate three results: |
|  |

|  |
| --- |
| \begin{itemize} |
|  |

|  |
| --- |
| \item There is statistical evidence that the textual production of the Erd\"os sectors are different. |
|  |

|  |
| --- |
| This can be noticed from the high values of $c'$ on these tables, |
|  |

|  |
| --- |
| beyond reference values used for the acceptance of the |
|  |

|  |
| --- |
| null hypothesis. |
|  |

|  |
| --- |
| Also, we regarded as non-negligible |
|  |

|  |
| --- |
| the values often above 0.1 for the Kolmogorov-Smirnov statistic (the maximum difference between the cumulative distributions), |
|  |

|  |
| --- |
| which we recognized as relevant for assuming differences in the underlying distributions in our study (see Section~\ref{sec:ks}). |
|  |

|  |
| --- |
| \item Intermediary sectors sometimes exhibit greater differences |
|  |

|  |
| --- |
| from periphery and hubs than these extreme sectors between themselves |
|  |

|  |
| --- |
| (Tables~\ref{tab:kolTok} and~\ref{tab:kolSub}). |
|  |

|  |
| --- |
| This differentiation of the three sectors is an |
|  |

|  |
| --- |
| indicative that the Erd\"os Sectioning |
|  |

|  |
| --- |
| described in Section~\ref{sectioning} reveals meaningful |
|  |

|  |
| --- |
| sectors of the networks. |
|  |

|  |
| --- |
| \item Evidence of differences between sectors on the same network |
|  |

|  |
| --- |
| (Tables~\ref{tab:kolTok}-\ref{tab:kolPct}) is often greater than differences between the same sector from distinct lists (Tables~\ref{tab:kolSubInter}-\ref{tab:kolPctInter}). |
|  |

|  |
| --- |
| \end{itemize} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \input{tables/ksTokens\_} |
|  |

|  |
| --- |
| \input{tables/ksWords\_} |
|  |

|  |
| --- |
| \input{tables/ksSents\_} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \input{tables/ksAdjs\_} |
|  |

|  |
| --- |
| \input{tables/ksSubs\_} |
|  |

|  |
| --- |
| \input{tables/ksPuns\_} |
|  |

|  |
| --- |
| \input{tables/ksInter} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| % We can summarize these results stating that the difference |
|  |

|  |
| --- |
| % found between texts from distinct Erd\"os sectors |
|  |

|  |
| --- |
| % is further evidenced. |
|  |

|  |
| --- |
| % We can summarize these results stating that the difference |
|  |

|  |
| --- |
| % found between the texts produced by the Erd\"os sectors |
|  |

|  |
| --- |
| % are greater than the difference found between texts from different |
|  |

|  |
| --- |
| % networks or from the same sector of different networks. |
|  |

|  |
| --- |
| \FloatBarrier |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Characters}\label{sec:cha} |
|  |

|  |
| --- |
| \input{tables/charsInline\_} |
|  |

|  |
| --- |
| Most often, peripheral and intermediary sectors use more digits and upper case letters. |
|  |

|  |
| --- |
| Hubs use more letters and vowels among letters. |
|  |

|  |
| --- |
| The use of white spaces, for example, does not seem to have any relation to connectivity. |
|  |

|  |
| --- |
| These results are illustrated in Table~\ref{tab:cha}. |
|  |

|  |
| --- |
| % The total number of characters in ELE list, |
|  |

|  |
| --- |
| % in the 20 thousand messages, |
|  |

|  |
| --- |
| % is more than three times what other lists exhibited. |
|  |

|  |
| --- |
| % This suggests peculiarities related to communication conventions and style (see Appendix~\ref{sec:materials}) and were found not related to topological features. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Tokens and words}\label{subsec:tw} |
|  |

|  |
| --- |
| %\input{tables/tokensInline\_} |
|  |

|  |
| --- |
| \input{tables/tokensMergedInline} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| %The largest average size of tokens is from the most wordy list (ELE). |
|  |

|  |
| --- |
| %This implies that is has more characters, tokens, and characters per token in comparison to the other lists. |
|  |

|  |
| --- |
| In most of the networks analyzed, hubs use longer words, which might be related to the use of a specialized vocabulary. |
|  |

|  |
| --- |
| Hubs use more contractions and known words, while peripheral sector exhibit a greater incidence of punctuations among tokens. |
|  |

|  |
| --- |
| Although the token diversity ($\frac{|tokens \neq|}{|tokens|}$) found in peripheral sector is far greater, |
|  |

|  |
| --- |
| this result has the masking artifact that the peripheral sector corpus is smaller, yielding a larger token diversity. |
|  |

|  |
| --- |
| This can be noticed by the token diversity of the whole network, which is lower than in any of the sectors. |
|  |

|  |
| --- |
| The same observation apply to the lexical diversity ($\frac{|kw\neq|}{kw}$). |
|  |

|  |
| --- |
| These results are exemplified in Table~\ref{tab:tokensInline}. |
|  |

|  |
| --- |
| % medir para tamanhos equivalentes de texto TTM |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| % ELE and CPP both exhibit intermediaries |
|  |

|  |
| --- |
| % with the more frequent production of punctuation, |
|  |

|  |
| --- |
| % less frequent production of known words, and |
|  |

|  |
| --- |
| % the highest incidence of words with wordnet synsets among known words. |
|  |

|  |
| --- |
| % This suggests some peculiarity in network structure, |
|  |

|  |
| --- |
| % such as authorities in the intermediary sector of such networks, |
|  |

|  |
| --- |
| % using smaller sentences and a more intensive use of jargons, |
|  |

|  |
| --- |
| % as made explicit in the following sections. |
|  |

|  |
| --- |
| % verificar isso e ver o que fazer |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| % Words with synsets, |
|  |

|  |
| --- |
| % among known English words, are less frequent in hubs sector, |
|  |

|  |
| --- |
| % evidencing the jargon and specialization of hubs. |
|  |

|  |
| --- |
| % verificar, desenvolvar |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| %\input{tables/tokenSizesInline\_} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Sizes of sentences}\label{subsec:ss} |
|  |

|  |
| --- |
| \input{tables/sentencesInline\_} |
|  |

|  |
| --- |
| Hubs present the lowest average sentence size, |
|  |

|  |
| --- |
| in characters, tokens or known words. |
|  |

|  |
| --- |
| This result is illustrated in Table~\ref{tab:sizesSents} |
|  |

|  |
| --- |
| and might be considered counterintuitive given that punctuation |
|  |

|  |
| --- |
| is more abundant in the texts of less connected participants. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Messages}\label{subsec:mm} |
|  |

|  |
| --- |
| \input{tables/messagesInline\_} |
|  |

|  |
| --- |
| Connectivity was found correlated to smaller messages in terms of characters and tokens. |
|  |

|  |
| --- |
| Connectivity was also found correlated to smaller messages in terms of the number of sentences, but |
|  |

|  |
| --- |
| it was less consistent. |
|  |

|  |
| --- |
| This result is exemplified in Table~\ref{tab:sizesMsgs}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| % tirar medidas de correlacao: perason e spearmanA TTM |
|  |

|  |
| --- |
| % ELE list displayed an inverse situation: |
|  |

|  |
| --- |
| % the more connected the sector, |
|  |

|  |
| --- |
| % the longer the messages are. |
|  |

|  |
| --- |
| % This was considered a peculiarity of the culture bonded |
|  |

|  |
| --- |
| % with the political subject of ELE list, to be further verified. |
|  |

|  |
| --- |
| % make some sort of verification TTM |
|  |

|  |
| --- |
| % process the lists just for this feature |
|  |

|  |
| --- |
| % ver a correlação das medidas de grau e força com TTM |
|  |

|  |
| --- |
| % texto |
|  |

|  |
| --- |
| % ver em especial se algo eh mais para in ou out |
|  |

|  |
| --- |
| % e se indifere greu e força como no particionamento de erdos |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{POS tags}\label{subsec:pos} |
|  |

|  |
| --- |
| \input{tables/posInline\_} |
|  |

|  |
| --- |
| % rever taggeamento e tentar melhor accuracy com o Bril Tagger TTM |
|  |

|  |
| --- |
| We found that lower connectivity yields more nouns and less verbs and adverbs. |
|  |

|  |
| --- |
| Also, the fraction of adjectives does not change, |
|  |

|  |
| --- |
| but given that peripherals use more nouns, |
|  |

|  |
| --- |
| we can conclude that hubs use more adjectives per noun. |
|  |

|  |
| --- |
| This suggests that the networks gather issues |
|  |

|  |
| --- |
| through the peripheral sector. |
|  |

|  |
| --- |
| These issues are qualified and proposed to be acted upon |
|  |

|  |
| --- |
| by the more connected participants. |
|  |

|  |
| --- |
| This is a further indicative that peripheral sectors |
|  |

|  |
| --- |
| are related to diversity while hubs relate to specialization. |
|  |

|  |
| --- |
| These results are exemplified in Table~\ref{tab:pos}. |
|  |

|  |
| --- |
| Weaker evidence was found that hubs use more \emph{adpositions}, |
|  |

|  |
| --- |
| determinants and 'particles and other functional words' while |
|  |

|  |
| --- |
| peripherals use more numerals. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Wordnet-related results} |
|  |

|  |
| --- |
| For correctly analyzing text production in terms of the Wordnet lexical database, |
|  |

|  |
| --- |
| we only considered words that had synsets of with the POS tag obtained with the POS tagger. |
|  |

|  |
| --- |
| This resulted in portions of tokens considered of $\approx 30\%$, |
|  |

|  |
| --- |
| but of more than $90\%$ of all tokens with Wordnet synsets. |
|  |

|  |
| --- |
| This yields less strong results, but which we found still relevant. |
|  |

|  |
| --- |
| Measures regarding Wordnet synsets often reach an extreme value (maximum or minimum) |
|  |

|  |
| --- |
| in the intermediary sector, which we understood as evidence that: |
|  |

|  |
| --- |
| \begin{itemize} |
|  |

|  |
| --- |
| \item the Erd\"os sectioning of the networks into peripheral, intermediary and hubs sectors are in fact relevant for human social structures, at least to the ones analyzed in this thesis. |
|  |

|  |
| --- |
| \item Human social networks present relations between connectivity and semantics. |
|  |

|  |
| --- |
| \item The intermediary sector might hold a deeper identity than that of a sector bounded by hubs and periphery sectors. |
|  |

|  |
| --- |
| \item The analysis of social networks texts using Wordnet reveals aspects of the structures which are not clear with the non-semantic analysis we performed. |
|  |

|  |
| --- |
| \end{itemize} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Furthermore, the analysis of the measures we obtained by means of the Wordnet |
|  |

|  |
| --- |
| is not trivial because of the number of different measures |
|  |

|  |
| --- |
| and because differences in measures are not obviously relevant. |
|  |

|  |
| --- |
| In order to obtain consistent results, we considered \emph{weak evidence of difference} in sectors in a network |
|  |

|  |
| --- |
| if maximum measure is at least 10\% greater than minimum measure, |
|  |

|  |
| --- |
| i.e. $\frac{maximum\;measure}{minumum\;measure}>1.1$. |
|  |

|  |
| --- |
| We considered \emph{evidence of difference} in sectors in a network if |
|  |

|  |
| --- |
| $\frac{maximum\;measure}{minumum\;measure}>1.2$. |
|  |

|  |
| --- |
| When |
|  |

|  |
| --- |
| $\frac{maximum\;measure}{minumum\;measure}>1.5$, we considered \emph{strong evidence of difference}. |
|  |

|  |
| --- |
| We than looked through each measure in all networks to reach compelling observations about the |
|  |

|  |
| --- |
| differences of sectors through all networks. |
|  |

|  |
| --- |
| Also useful here is the definition of lower sectors (peripheral and intermediary), |
|  |

|  |
| --- |
| upper sectors (intermediary and hubs) and extreme sectors (peripheral and hubs). |
|  |

|  |
| --- |
| We should also point when measurements peak at the intermediary sector, |
|  |

|  |
| --- |
| be it a maximum or minimum peak. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Noteworthy is that extra skepticism should be kept in mind about these results |
|  |

|  |
| --- |
| because of the unquantified noise in the measurements. |
|  |

|  |
| --- |
| Observations seem consistent and meaningful, but only about a third of total tokens |
|  |

|  |
| --- |
| were considered. |
|  |

|  |
| --- |
| This is because tokens were discarded if not having a Wordnet synset |
|  |

|  |
| --- |
| or when not having a synset with a POS tag true to the POS tag attributed by the POS tagger. |
|  |

|  |
| --- |
| Besides that, there are often more than one synset with the same POS tag for each word, |
|  |

|  |
| --- |
| and we chose the most frequent synset as ranked by Wordnet. |
|  |

|  |
| --- |
| In the positive side, we observe that $\approx 95\%$ of tokens which had synsets |
|  |

|  |
| --- |
| were considered. |
|  |

|  |
| --- |
| Example of types tokens without synsets are stopwords, punctuations, numerals, acronyms and typos. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsubsection{Wordnet POS tags}\label{subsec:wnpos} |
|  |

|  |
| --- |
| The observations here are somewhat consistent with those in Section~\ref{subsec:pos}: |
|  |

|  |
| --- |
| peripherals use more nouns and less verbs and adverbs. |
|  |

|  |
| --- |
| The difference here is regarding adjectives, which was found more frequent in hubs texts |
|  |

|  |
| --- |
| in this reduced set of tokens. |
|  |

|  |
| --- |
| These results are illustrated in Table~\ref{tab:wnpos}. |
|  |

|  |
| --- |
| \input{tables/wnPOSInline\_} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsubsection{Wordnet synsets characteristics}\label{subsec:wn1} |
|  |

|  |
| --- |
| Wordnet synsets with different POS tags have different relations. |
|  |

|  |
| --- |
| Therefore, we made separate observations about each POS tag: |
|  |

|  |
| --- |
| \begin{itemize} |
|  |

|  |
| --- |
| \item Nouns, exemplified in Table~\ref{tab:wnn}. |
|  |

|  |
| --- |
| \input{tables/wnPOSInline2-n-13tag\_} |
|  |

|  |
| --- |
| \begin{itemize} |
|  |

|  |
| --- |
| \item Minimum and maximum depth: |
|  |

|  |
| --- |
| differences were found in the mean of minimum and maximum depth of a synset between email lists, |
|  |

|  |
| --- |
| but not once among sectors of a network. |
|  |

|  |
| --- |
| Differences between the variance of minimum and maximum depth of synsets of sectors was found mostly nonexistent or weak. |
|  |

|  |
| --- |
| \item Holonyms: |
|  |

|  |
| --- |
| differences in the number of holonyms per word were present in $\approx 85\%$ of the networks and were |
|  |

|  |
| --- |
| more incident in the lower sectors in $\approx 90\%$ of the observations in which we found such differences. |
|  |

|  |
| --- |
| Differences in the variance in the number of holonyms was also found with the same regularity, |
|  |

|  |
| --- |
| but were greater in the upper sectors in $\approx 80\%$ of the networks. |
|  |

|  |
| --- |
| Both mean and variance of the number of holonyms peaked in the intermediary sector in $\approx50\%$ of the observations. |
|  |

|  |
| --- |
| \item Meronyms: |
|  |

|  |
| --- |
| words with more meronyms were present in $\approx 90\%$ of the networks and were |
|  |

|  |
| --- |
| more incident in the lower sectors in $\approx 80\%$ of the observations in which we found such differences. |
|  |

|  |
| --- |
| Differences in the variance in the number of meronyms was found in $100\%$ of the networks and was often strong. |
|  |

|  |
| --- |
| The variance was greater in the periphery in $66.66\%$ and in the lower sectors in $\approx 90\%$ of the observations. |
|  |

|  |
| --- |
| \item Domain: |
|  |

|  |
| --- |
| differences in the mean and variance of the number of domains of words were found respectively in $90\%$ and $50\%$ of the networks and maximum values were found evenly distributed across sectors. |
|  |

|  |
| --- |
| Peaks were found in the intermediary sector in $\approx 50\%$ of the networks. |
|  |

|  |
| --- |
| \item Lemmas: |
|  |

|  |
| --- |
| differences in the mean and variance of the number of lemmas of words were found respectively in $40\%$ and $55\%$ of the networks. |
|  |

|  |
| --- |
| In $\approx 90\%$ of the cases where there was difference in the mean, the maximum number of lemmas was found in the periphery. |
|  |

|  |
| --- |
| Peaks in the intermediary sector were less often, occurring only in $\approx 35\%$ of the observations. |
|  |

|  |
| --- |
| \item Hyponyms: |
|  |

|  |
| --- |
| differences in the mean and variance of the number of hyponyms of words were found respectively in $77.77\%$ and $88.88\%$ of the networks. |
|  |

|  |
| --- |
| In $\approx 93\%$ of the cases where there was difference in the mean, |
|  |

|  |
| --- |
| the maximum number of hyponyms was found indistinctly in the upper sectors. |
|  |

|  |
| --- |
| In $75\%$ of the cases where there was difference in the variance, |
|  |

|  |
| --- |
| the maximum variance was found indistinctly in the upper sectors. |
|  |

|  |
| --- |
| Peaks occurred for both mean and variance in the intermediary sector in $\approx 75\%$ of the observations. |
|  |

|  |
| --- |
| \item Hypernyms: |
|  |

|  |
| --- |
| between the sectors of all networks analyzed, we found no differences in the mean of the number of hypernyms. |
|  |

|  |
| --- |
| There were differences in the variance of the number of hypernyms of the words used by the sectors in $\approx 72\%$ of the networks. |
|  |

|  |
| --- |
| Greatest values occurred indistinctly in all sectors and peaked in the intermediary sector in $\approx 50\%$ of the observations. |
|  |

|  |
| --- |
| \end{itemize} |
|  |

|  |
| --- |
| \item Adjectives, exemplified in Table~\ref{tab:wnas}. |
|  |

|  |
| --- |
| \input{tables/wnPOSInline2-as-9tag\_} |
|  |

|  |
| --- |
| \begin{itemize} |
|  |

|  |
| --- |
| \item Domain: |
|  |

|  |
| --- |
| differences in the mean and variance of the number of domains of words were found respectively in $88.88\%$ and $61.11\%$ of the networks. |
|  |

|  |
| --- |
| In $87.5\%$ of the cases where there was difference in the mean, |
|  |

|  |
| --- |
| the maximum number of domains was found indistinctly in the upper sectors. |
|  |

|  |
| --- |
| In $\approx 82\%$ of the cases where there was difference in the variance, |
|  |

|  |
| --- |
| the maximum variance was found indistinctly in the upper sectors. |
|  |

|  |
| --- |
| Peaks occurred in the intermediary sector in $68.75\%$ of the observations for the mean |
|  |

|  |
| --- |
| and in $\approx 54.55\%$ of the observations for the variance. |
|  |

|  |
| --- |
| \item Similar: |
|  |

|  |
| --- |
| differences in the mean and variance of the number of similar synsets relations of adjectives were found respectively only in $44.45\%$ and $61.11\%$ of the networks. |
|  |

|  |
| --- |
| In $\approx 90\%$ of the cases where there was difference in the mean, |
|  |

|  |
| --- |
| the maximum number of domains was found in the hubs sector. |
|  |

|  |
| --- |
| In $\approx 90\%$ of the cases where there was difference in the variance, |
|  |

|  |
| --- |
| the maximum number of domains was found indistinctly the extreme sectors. |
|  |

|  |
| --- |
| Peaks occurred in the intermediary sector in $50\%$ of the observations for the mean |
|  |

|  |
| --- |
| and in $\approx 36.37\%$ of the observations for the variance. |
|  |

|  |
| --- |
| \item Lemmas: |
|  |

|  |
| --- |
| differences in the mean and variance of the number of lemmas of adjectives were found respectively only in $27.78\%$ and $72.22\%$ of the networks. |
|  |

|  |
| --- |
| Maximum values occurred indistinctly in all sectors and peaks were found in the intermediary sector in $\approx 50\%$ of the observed cases. |
|  |

|  |
| --- |
| \end{itemize} |
|  |

|  |
| --- |
| \item Verbs, illustrated in Table~\ref{tab:wnv}. |
|  |

|  |
| --- |
| \input{tables/wnPOSInline2-v-3tag\_} |
|  |

|  |
| --- |
| \begin{itemize} |
|  |

|  |
| --- |
| \item No significant differences were found in the mean and variance verb synset relations of minimum and maximum depth, verb groups, lemmas and hypernyms. |
|  |

|  |
| --- |
| \item Domains and entailments: |
|  |

|  |
| --- |
| differences were often strong (i.e. $>1.5$) in both mean and variance. |
|  |

|  |
| --- |
| Due to the reduced number of verbs and the small values of mean and variance, |
|  |

|  |
| --- |
| we considered these measures as not significant. |
|  |

|  |
| --- |
| \item Hyponyms: |
|  |

|  |
| --- |
| differences in the mean and variance of the number of hyponyms of verbs were found respectively in $50\%$ and $72.23\%$ of the networks. |
|  |

|  |
| --- |
| In $\approx 90\%$ of the cases where there was difference in the mean, |
|  |

|  |
| --- |
| the maximum number of hyponyms was found in the upper sectors ($66.67\%$ in the hubs sector). |
|  |

|  |
| --- |
| In $\approx 85\%$ of the cases where there was difference in the variance, |
|  |

|  |
| --- |
| the maximum number of domains was found indistinctly the upper sectors ($61.54\%$ in the hubs sector). |
|  |

|  |
| --- |
| Peaks occurred in the intermediary sector in $\approx 35\%$ with respect to both mean and variance. |
|  |

|  |
| --- |
| \end{itemize} |
|  |

|  |
| --- |
| \item Adverbs, exemplified in Table~\ref{tab:wnr}. |
|  |

|  |
| --- |
| \input{tables/wnPOSInline2-r-17tag\_} |
|  |

|  |
| --- |
| \begin{itemize} |
|  |

|  |
| --- |
| \item Domains: |
|  |

|  |
| --- |
| differences in the mean and variance of the number of domains of adverbs were found respectively in $\approx 95.45\%$ and $\approx 66.67\%$ of the networks. |
|  |

|  |
| --- |
| In $\approx 82.35\%$ of the cases where there was difference in the mean, |
|  |

|  |
| --- |
| the maximum number of domains was found in the upper sectors ($58.82\%$ in the hubs sector). |
|  |

|  |
| --- |
| In $\approx 92\%$ of the cases where there was difference in the variance, |
|  |

|  |
| --- |
| the maximum number of domains was found indistinctly the upper sectors ($50\%$ in the intermediary sector). |
|  |

|  |
| --- |
| Peaks occurred in the intermediary sector in $\approx 64.71\%$ and $75\%$ in the mean and variance respectively. |
|  |

|  |
| --- |
| \item Lemmas: |
|  |

|  |
| --- |
| no systematic difference was found in the mean and variance of the number of lemmas of adverbs. |
|  |

|  |
| --- |
| \end{itemize} |
|  |

|  |
| --- |
| \end{itemize} |
|  |

|  |
| --- |
| \subsubsection{Wordnet synset hypernyms}\label{subsec:wn1} |
|  |

|  |
| --- |
| \begin{itemize} |
|  |

|  |
| --- |
| \item Noun synsets: differences in the use of nouns with physical entity hypernyms was found indistincly in all sectors. |
|  |

|  |
| --- |
| In deeper layers, more systematic differences arises. |
|  |

|  |
| --- |
| With depth 2, hubs use more nouns related to attribute and psycological features. |
|  |

|  |
| --- |
| Lower sectors use more nouns related to measure, |
|  |

|  |
| --- |
| with $62.5\%$ of the lists where this difference was found having greater values in the peripheral sector. |
|  |

|  |
| --- |
| Communication related nouns were found mostly in extreme sectors. |
|  |

|  |
| --- |
| With depth 3, hubs presented more nouns related to written communication, event and cognition. |
|  |

|  |
| --- |
| Peripherals showed greater use of nouns related to definite quantity. |
|  |

|  |
| --- |
| Message related nouns often peaked at the intermediary sector. |
|  |

|  |
| --- |
| These results are shown in Table~\ref{tab:wnnh}. |
|  |

|  |
| --- |
| \input{tables/finalSynsetsN\_} % uma soh |
|  |

|  |
| --- |
| \item Adjective synsets: the use of adjectives was found less systematic. |
|  |

|  |
| --- |
| The synsets varied greatly among lists and differences were not strong. |
|  |

|  |
| --- |
| We obseved weak evidence that hubs use more adjectives related to certainty, |
|  |

|  |
| --- |
| and that the use of such adjectives always peaked at the intermediary sector. |
|  |

|  |
| --- |
| Even weaker evidence was found that hubs use more adjectives related to newness. |
|  |

|  |
| --- |
| These results are shown in Table~\ref{tab:wnash}. |
|  |

|  |
| --- |
| \input{tables/finalSynsetsAS\_} % uma soh |
|  |

|  |
| --- |
| \item Verbs synset hypernyms of move and travel was found more numerous in the peripheral sector. |
|  |

|  |
| --- |
| Verbs related to change was found more common in the hubs sector. |
|  |

|  |
| --- |
| Verbs related to making was found with differences in the frequency of use among sectors, |
|  |

|  |
| --- |
| but had greatest incidence in all sectors (but in distinct networks). |
|  |

|  |
| --- |
| With depth 2, hubs exhibited greater use of verbs related to state and evaluate while |
|  |

|  |
| --- |
| peripherals exhibited greater use of verbs related to keeping and putting. |
|  |

|  |
| --- |
| With depth 3, in the upper sectors was found a greater use of verbs related to thinking. |
|  |

|  |
| --- |
| Hubs used more increase-related verbs. |
|  |

|  |
| --- |
| Periphery presented more verbs related to running and communication. |
|  |

|  |
| --- |
| With depth 4, lower sectors used more verbs related to informing, peripherals might be regarded as using more verbs related to recording (set in a permanent form), and hubs as using more verbs related to adding. |
|  |

|  |
| --- |
| These results are shown in Table~\ref{tab:wnvh}. |
|  |

|  |
| --- |
| \input{tables/finalSynsetsV\_} % uma soh |
|  |

|  |
| --- |
| \item Adverb synsets was found with particularly interesting patterns as greater use of adverbs related to possibility and stillness was found in the intermediary sector. |
|  |

|  |
| --- |
| Adverbs related to however and even were more frequent in the peripheral sector while adverbs related to well (good way to perform) was more used by hubs. |
|  |

|  |
| --- |
| These results are shown in Table~\ref{tab:wnrh}. |
|  |

|  |
| --- |
| \input{tables/finalSynsetsR\_} % uma soh |
|  |

|  |
| --- |
| \end{itemize} |
|  |

|  |
| --- |
| % \input{tables/wnPOSInline2} % uma soh |
|  |

|  |
| --- |
| %\input{tables/wnInline\_} % uma soh |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| % \subsection{Wordnet synsets}\label{subsec:wnres} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| %\input{tables/wnInline2-n\_} % uma para cada nivel do synset |
|  |

|  |
| --- |
| %\input{tables/wnInline2a-n\_} |
|  |

|  |
| --- |
| %\input{tables/wnInline2b-n\_} |
|  |

|  |
| --- |
| %\input{tables/wnInline2c-n\_} |
|  |

|  |
| --- |
| %\input{tables/wnInline2d-n\_} |
|  |

|  |
| --- |
| % |
|  |

|  |
| --- |
| %\input{tables/wnInline2-as\_} % uma para cada nivel do synset |
|  |

|  |
| --- |
| %\input{tables/wnInline2a-as\_} |
|  |

|  |
| --- |
| %%\input{tables/wnInline2b-as\_} |
|  |

|  |
| --- |
| %%\input{tables/wnInline2c-as\_} |
|  |

|  |
| --- |
| %%\input{tables/wnInline2d-as\_} |
|  |

|  |
| --- |
| % |
|  |

|  |
| --- |
| %\input{tables/wnInline2-v\_} % uma para cada nivel do synset |
|  |

|  |
| --- |
| %\input{tables/wnInline2a-v\_} |
|  |

|  |
| --- |
| %\input{tables/wnInline2b-v\_} |
|  |

|  |
| --- |
| %\input{tables/wnInline2c-v\_} |
|  |

|  |
| --- |
| %\input{tables/wnInline2d-v\_} |
|  |

|  |
| --- |
| % |
|  |

|  |
| --- |
| %\input{tables/wnInline2-r\_} % uma para cada nivel do synset |
|  |

|  |
| --- |
| %\input{tables/wnInline2a-r\_} |
|  |

|  |
| --- |
| %%\input{tables/wnInline2b-r\_} |
|  |

|  |
| --- |
| %%\input{tables/wnInline2c-r\_} |
|  |

|  |
| --- |
| %%\input{tables/wnInline2d-r\_} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Correlation of topological and textual metrics}\label{subsec:cor} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \input{tables/correlationInline9\_} |
|  |

|  |
| --- |
| Overall, small correlation is found between textual and topological metrics. |
|  |

|  |
| --- |
| An exception is that, in the hubs sector, strengh was very often negatevely correlated to the mean and variance of the number |
|  |

|  |
| --- |
| of punctuations (and sometimes with the number of known words or stopwords) |
|  |

|  |
| --- |
| with values bellow -0.4, but a few positive and high values (above 0.5) were also found. |
|  |

|  |
| --- |
| % verificar e quantificar TTM |
|  |

|  |
| --- |
| Interestingly, the number of punctuations per sentence was most often correlated to the number of stopwords |
|  |

|  |
| --- |
| while most often \emph{not} correlated to the number of known words. |
|  |

|  |
| --- |
| Noteworthy is that degree is negatively correlated to clustering coefficient in intermediary and hubs sectors, |
|  |

|  |
| --- |
| which is consistent with the literature, but it is positively correlated for |
|  |

|  |
| --- |
| peripheral sectors. |
|  |

|  |
| --- |
| Other strong correlation associations of textual and topological measures were found |
|  |

|  |
| --- |
| but not systematically and might be indicatives of style from the different lists analyzed. |
|  |

|  |
| --- |
| These results are exemplified in Table~\ref{tab:cor}. |
|  |

|  |
| --- |
| % verificar e fazer tabelas corretamente: |
|  |

|  |
| --- |
| % focadas para texto, extensivas para SI |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Formation of principal components}\label{subsec:pc} |
|  |

|  |
| --- |
| \input{tables/pcaInline11\_} |
|  |

|  |
| --- |
| % \input{tables/pcaInline\_} |
|  |

|  |
| --- |
| % \begin{figure}[!h] |
|  |

|  |
| --- |
| % \centering |
|  |

|  |
| --- |
| % \caption{First two principal components.} |
|  |

|  |
| --- |
| % \includegraphics[width=0.3\textwidth]{figs/plot\_pca} |
|  |

|  |
| --- |
| % \label{fig:formation} |
|  |

|  |
| --- |
| % \end{figure} |
|  |

|  |
| --- |
| Principal components formation of textual and topological metrics |
|  |

|  |
| --- |
| seems to be the less stable of all results reported in this study. |
|  |

|  |
| --- |
| The concentration of dispersion often peaked in the intermediary sector. |
|  |

|  |
| --- |
| % Clustering coefficient was dominant in a principal component almost exclusively |
|  |

|  |
| --- |
| % when the whole network was used for PCA, otherwise it combined more evenly with |
|  |

|  |
| --- |
| % other measures. |
|  |

|  |
| --- |
| % This reveals that clustering coefficient might be a more relevant feature for the complete network |
|  |

|  |
| --- |
| % than for each separate sector. |
|  |

|  |
| --- |
| Components are most often composed of topological or textual features. |
|  |

|  |
| --- |
| Other than that, we observe that PCA is sensitive to metrics included |
|  |

|  |
| --- |
| and should reveal other insights in other settings. |
|  |

|  |
| --- |
| These results are exemplified in Table~\ref{tab:tpca}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Results still to be interpreted}\label{subsec:sii} |
|  |

|  |
| --- |
| %These networks yield diverse characteristics, |
|  |

|  |
| --- |
| %some of which were not of core importance for this step of the research. |
|  |

|  |
| --- |
| %Even so, at least one of these characteristics was found interesting enough to be considered a result and an example of interesting artifacts found. |
|  |

|  |
| --- |
| Histogram differences of incident word sizes with and without repetition of words are constant. |
|  |

|  |
| --- |
| That is, in each email list, when a histogram of word sizes were made with all words written, |
|  |

|  |
| --- |
| and another histogram made with sizes of all \emph{different} words, |
|  |

|  |
| --- |
| the cumulative absolute difference of the two histograms throughout the bins were found constant for all lists analysed. |
|  |

|  |
| --- |
| When all known English words were considered, |
|  |

|  |
| --- |
| the difference sums up to $\approx 1.0$. |
|  |

|  |
| --- |
| When stopwords are discarded, |
|  |

|  |
| --- |
| the difference found was different, but still constant, slightly above $0.5$. |
|  |

|  |
| --- |
| When only stopwords were considered, the difference is $\approx 0.6$. |
|  |

|  |
| --- |
| When only known English words that does not have Wordnet synsets are used, |
|  |

|  |
| --- |
| this difference is $\approx 1.2$. |
|  |

|  |
| --- |
| We considered this result a number of times in the past years and presented it to other researchers, |
|  |

|  |
| --- |
| but reached no conclusions about its meaning. |
|  |

|  |
| --- |
| Appendix ~\ref{sec:resE} are dedicated to this histogram differences. |
| Comecei a revisão a partir daqui |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \section{Results from visualization} |
|  |

|  |
| --- |
| The results associated with Versinus are divided into two groups: |
|  |

|  |
| --- |
| observations on features that made it useful for the task of analyzing the general properties of human interaction networks, |
|  |

|  |
| --- |
| and the network properties it made possible to grasp. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Useful visualization features for dynamic networks} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Among the numerous insights related to Versinus, a few seem more fundamental, while others were simply useful. These |
|  |

|  |
| --- |
| insights were incorporated to Versinus as the result of tests which presented clear benefits within the context of our research. |
|  |

|  |
| --- |
| The following list is an attempt to present the observations and insights in an importance-first order: |
|  |

|  |
| --- |
| \begin{enumerate} |
|  |

|  |
| --- |
| \item Vertices need to remain static. |
|  |

|  |
| --- |
| Even if they move smoothly, one should notice solely transient artifacts from the structure. |
|  |

|  |
| --- |
| \item Very connected sectors (hubs and intermediary) need to be in a curve, otherwise the edges enclose each other and it is difficult to infer information from the network. |
|  |

|  |
| --- |
| \item The height and width of a vertex are very informative, especially if metrics mapped to them have a strong relation, such as out-degree (mapped to height in Versinus) and in-degree (mapped to width). |
|  |

|  |
| --- |
| \item The color of nodes is also informative although less than height and weight, as differences in the latter are more noticeable. |
|  |

|  |
| --- |
| \item An ordering of nodes, related to their fixed position, is very useful. Among all tests, ordering of vertices by degree was considered the most informative, which led to the hub, intermediary and peripheral sectioning of the network delineated in Section~\ref{sectioning}. |
|  |

|  |
| --- |
| As node position in the layout is fixed throughout an animation which comprises consecutive but distinct network activity, such ordering is done with respect to the resulting network of all the activity. |
|  |

|  |
| --- |
| Numbering these positions with respect to the order of the vertices in the larger structure (i.e. all $M$ messages) is useful for understanding to which extent a vertex preserves the position in different scales of activity. |
|  |

|  |
| --- |
| \end{enumerate} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Many other insights were derived from Versinus, such as possible visualization tools, other kinds of convenient layouts and glyph elaborations. |
|  |

|  |
| --- |
| These received attention in Section~\ref{sec:verref}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Understanding of network properties through Versinus} |
|  |

|  |
| --- |
| A number of hypotheses were drawn about the networks for which Versinus was designed. |
|  |

|  |
| --- |
| As suggested by Palla, Barab\'asi and Vicsek~\cite{barabasiEvo}, stability of participant activity in social networks is more incident in smaller networks. |
|  |

|  |
| --- |
| Consistent with this result, all hubs have intermittent activity in the settings analyzed, except for the email list with the smallest number of participants (the Metareciclagem email list). |
|  |

|  |
| --- |
| The intermittence of hubs was one of the top hypotheses which motivated the development of Versinus. |
|  |

|  |
| --- |
| The stability of the network structure, concomitant with the instability of the activity of each participant, |
|  |

|  |
| --- |
| motivated a deeper analysis~\cite{stab}. |
|  |

|  |
| --- |
| In doing so, we also found evidence for another hypothesis drawn from Versinus: |
|  |

|  |
| --- |
| that in- and out-degree differences in each vertex are important for network characterization. |
|  |

|  |
| --- |
| Furthermore, the visualization suggests that there are modes of operation of the network. |
|  |

|  |
| --- |
| As an example, the intermediary sector often communicates mostly with the hubs or with the peripheral vertices. |
|  |

|  |
| --- |
| Other hypotheses, |
|  |

|  |
| --- |
| such as discrepancies in the authority and the degree of a vertex, |
|  |

|  |
| --- |
| are numerous but need further research to be validated. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Refinement of Versinus}\label{sec:verref} |
|  |

|  |
| --- |
| Versinus was convenient for obtaining insights about how to enhance its layout and use. |
|  |

|  |
| --- |
| It was immediate to think of a tool for using Versinus |
|  |

|  |
| --- |
| in real-time, but less obvious are some ideas about the layout and visual guides. |
|  |

|  |
| --- |
| To further enable visualization of hubs and intermediary vertex, |
|  |

|  |
| --- |
| the sinusoid can have many periods |
|  |

|  |
| --- |
| with a decaying frequency. |
|  |

|  |
| --- |
| The upper straight line can also have an oscillating outline. |
|  |

|  |
| --- |
| The two halves of the sinusoidal period could be moved independently. |
|  |

|  |
| --- |
| The waveform need not to be a sinusoid. |
|  |

|  |
| --- |
| One can think of many ways to make more informative glyphs. |
|  |

|  |
| --- |
| Also, visual and auditory signals for specific occurrences can be interesting |
|  |

|  |
| --- |
| (e.g. when a new vertex appears, when one vanishes, when an ordering of vertices changes). |
|  |

|  |
| --- |
| Measurements of each vertex can be shown with a vertical displacement, |
|  |

|  |
| --- |
| to enable multiple measurements, to avoid the need to blink the numbers and to keep network visualization free from occlusion. |
|  |

|  |
| --- |
| Working with Versinus has also suggested other kinds of layout for vertices, |
|  |

|  |
| --- |
| especially geometric figures and iterative force-based methods for positioning vertices in a fixed layout. |
|  |

|  |
| --- |
| The traditional matrix representation of the graphs has been gazed upon as support to Versinus |
|  |

|  |
| --- |
| as has been some recent approaches to network visualization~\cite{Viz1}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \section{Linked data results} |
|  |

|  |
| --- |
| \label{outline} |
|  |

|  |
| --- |
| The current results include data selection and preparation for knowledge discovery. |
|  |

|  |
| --- |
| In this respect, the main result lies in the fact that data were made available, which enables benchmarking of scientific results |
|  |

|  |
| --- |
| and experimentations. |
|  |

|  |
| --- |
| Secondary results include data outline through figures and tables, |
|  |

|  |
| --- |
| software support and example SparQL queries. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Standardization} |
|  |

|  |
| --- |
| The data is embedded into standard URIs and triples, i.e. translated to RDF. |
|  |

|  |
| --- |
| URIs are built in the namespace \url{http://purl.org/socialparticipation/participationontology/} |
|  |

|  |
| --- |
| which are identified herein with the prefix \textttt{po:}. |
|  |

|  |
| --- |
| Classes and properties are built by adding a suffix to the root, |
|  |

|  |
| --- |
| as in the class \textttt{po:Participant} or in the property \textttt{po:text}. |
|  |

|  |
| --- |
| Classes have ``UpperCamelCase'' suffixes while properties have ``lowerCamelCase'' suffixes. |
|  |

|  |
| --- |
| All class instances, such as participants, messages, friendships and |
|  |

|  |
| --- |
| interactions, are linked to |
|  |

|  |
| --- |
| snapshots through the triple \textttt{<instance> po:snapshot <snapshot\\_uri>}. |
|  |

|  |
| --- |
| Message texts, including comments, are objects in the triple: \textttt{<message\\_id> po:text <message\\_text>}. |
|  |

|  |
| --- |
| Preprocessed texts are objects of triples: \textttt{<message\\_id> po:cleanText <message\\_text>}. |
|  |

|  |
| --- |
| More specialized predicates are used for delivering text when necessary, |
|  |

|  |
| --- |
| such as \textttt{po:htmlBodyText} and \textttt{po:cleanBodyText} used |
|  |

|  |
| --- |
| for ParticipaBR articles (instances of the class \textttt{po:Article}. |
|  |

|  |
| --- |
| A participant URI is unique throughout the provenance (e.g. the same for |
|  |

|  |
| --- |
| the same participant in all Twitter snapshots). |
|  |

|  |
| --- |
| To enable annotations which differ when the snapshot changes, |
|  |

|  |
| --- |
| \textttt{po:Observation} class instances are used in the triple |
|  |

|  |
| --- |
| \textttt{<participant\\_uri> po:observation <observation\\_uri>}. |
|  |

|  |
| --- |
| The observation instances are then linked to the snapshot and the |
|  |

|  |
| --- |
| data. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| Instances are built on top of the class they derive from plus a hashtag character, |
|  |

|  |
| --- |
| a provenance string (e.g. \textttt{facebook-legacy} or |
|  |

|  |
| --- |
| \textttt{participabr-legacy}) of the snapshot they refer to, and an identifier; |
|  |

|  |
| --- |
| i.e. \textttt{po:Participant\#<provenance-legacy>-<id>}. |
|  |

|  |
| --- |
| All snapshot URIs follow the formation rule: \textttt{po:<SnapshotProvenance>\#<snapshot\\_id>}. |
|  |

|  |
| --- |
| All snapshot ids follow the formation rule: \textttt{<platform>-legacy-<further\\_identifier>}; e.g. |
|  |

|  |
| --- |
| \textttt{irc-legacy-labmacambira} or |
|  |

|  |
| --- |
| \textttt{email-legacy-linux.audio.devel1-20000}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Data outline} |
|  |

|  |
| --- |
| \input{misc/overallText} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \input{tables/nsnapshots} |
|  |

|  |
| --- |
| % stats: |
|  |

|  |
| --- |
| %% number of triples |
|  |

|  |
| --- |
| %% number of edges |
|  |

|  |
| --- |
| %% number of chars |
|  |

|  |
| --- |
| %% number of users |
|  |

|  |
| --- |
| % diagrams in the supporting information file |
|  |

|  |
| --- |
| \subsection{Software tools} |
|  |

|  |
| --- |
| The database is released with software for rendering itself, analyses and |
|  |

|  |
| --- |
| multimedia artifacts. |
|  |

|  |
| --- |
| \subsubsection{Triplification routines} |
|  |

|  |
| --- |
| For each social platform there is a \emph{triplification} routine, |
|  |

|  |
| --- |
| i.e. a script for translating data to RDF. |
|  |

|  |
| --- |
| Original formats and further observations are presented in |
|  |

|  |
| --- |
| Table~\ref{tab:provenance}. |
|  |

|  |
| --- |
| \begin{table\*}[h!]\scriptsize |
|  |

|  |
| --- |
| \begin{center} |
|  |

|  |
| --- |
| \caption{Social platforms, original formats and further observations for |
|  |

|  |
| --- |
| the database.}\label{tab:provenance} |
|  |

|  |
| --- |
| \begin{tabular}{| l || p{3cm} | p{3cm} | c |}\hline |
|  |

|  |
| --- |
| \textbf{social platform} & \textbf{original format} & \textbf{further observations} & \textbf{toolbox} \\\hline\hline |
|  |

|  |
| --- |
| AA & MySQL and MongoDB databases; IRC text logs & donated by AA users & Participation~\cite{participation} \\\hline |
|  |

|  |
| --- |
| Cidade Democrática & MySQL database & donated by admins & Participation \\\hline |
|  |

|  |
| --- |
| Email & mbox & obtained through Gmane public database & Gmane~\cite{gmane} \\\hline |
|  |

|  |
| --- |
| Facebook & GDF, GML and TAB & obtained through Netvizz~\cite{netvizz} & Social~\cite{social} \\\hline |
|  |

|  |
| --- |
| IRC & plain text log & obtained through Supybot logging & Social \\\hline |
|  |

|  |
| --- |
| ParticipaBR & PostgreSQL database & donated by admins & Participation \\\hline |
|  |

|  |
| --- |
| Twitter & JSON & obtained through Twitter streaming API & Social \\\hline |
|  |

|  |
| --- |
| \end{tabular} |
|  |

|  |
| --- |
| \begin{flushleft} |
|  |

|  |
| --- |
| Source: Prepared by the author.\ |
|  |

|  |
| --- |
| \end{flushleft} |
|  |

|  |
| --- |
| \end{center} |
|  |

|  |
| --- |
| \end{table\*} |
|  |

|  |
| --- |
| \subsubsection{Topological and textual analysis}\label{ana} |
|  |

|  |
| --- |
| Routines are available for taking the topological and textual measurements from |
|  |

|  |
| --- |
| the database. |
|  |

|  |
| --- |
| Auxiliary routines, such as performing principal component analysis |
|  |

|  |
| --- |
| and taking Kolmogorov-Smirnov measurements, are available |
|  |

|  |
| --- |
| to facilitate pattern recognition. |
|  |

|  |
| --- |
| All the analysis routines used for this thesis are in these publicly accessible scripts. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsubsection{Multimedia rendering}\label{media} |
|  |

|  |
| --- |
| It is a core purpose of the framework to provide routines for rendering |
|  |

|  |
| --- |
| audiovisualizations of the data. |
|  |

|  |
| --- |
| Social structures are rendered into music, images and video animations |
|  |

|  |
| --- |
| through the Percolation toolbox~\cite{percolation} in association with |
|  |

|  |
| --- |
| the Music and Visuals toolboxes~\cite{music,visuals}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsubsection{Migration from deprecated toolboxes} |
|  |

|  |
| --- |
| Routines mentioned in Sections~\ref{ana} and~\ref{media} are being migrated from deprecated |
|  |

|  |
| --- |
| toolboxes~\cite{gmaneLegacy,percolationLegacy} into newly designed |
|  |

|  |
| --- |
| toolboxes~\cite{percolation,visuals}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Diagrams of the data and auxiliary tables} |
|  |

|  |
| --- |
| The database exploration can be assisted through diagrams which shows |
|  |

|  |
| --- |
| the structure from each provenance. |
|  |

|  |
| --- |
| Such diagrams are exemplified in the Appendix~\ref{losd} and |
|  |

|  |
| --- |
| are fully available in an article~\cite{losd} |
|  |

|  |
| --- |
| with some tables to make it easier to understand the data provided. |
|  |

|  |
| --- |
| A simplified example is given in Figure~\ref{dia} where the friendship |
|  |

|  |
| --- |
| structure of the Facebook snapshots is illustrated. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \begin{figure}[!ht] |
|  |

|  |
| --- |
| \centering |
|  |

|  |
| --- |
| \caption{A diagram of the structure involved in the friendship networks |
|  |

|  |
| --- |
| of the Facebook snapshots. |
|  |

|  |
| --- |
| A green edge denotes an OWL existential class restriction; |
|  |

|  |
| --- |
| an inverted nip denotes an OWL universal class restriction; |
|  |

|  |
| --- |
| a full (non-dashed) edge denotes an OWL functional property axiom. |
|  |

|  |
| --- |
| Further information and complete diagrams for each provenance are given in ~\cite{losd}.}\label{dia} |
|  |

|  |
| --- |
| \includegraphics[width=0.5\textwidth]{ontologies/facebook-legacy-AntonioAnzoategui18022013Friendship.ttl/draw} |
|  |

|  |
| --- |
| \begin{flushleft} |
|  |

|  |
| --- |
| Source: Prepared by the author.\ |
|  |

|  |
| --- |
| \end{flushleft} |
|  |

|  |
| --- |
| \end{figure} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{SPARQL queries}\label{queries} |
|  |

|  |
| --- |
| There are numerous useful and general purpose SPARQL queries to be performed against the database. |
|  |

|  |
| --- |
| Here we write some of such queries selected by their simplicity and potential to be varied. |
|  |

|  |
| --- |
| All queries assume the use of the preamble \textttt{PREFIX po: <http://purl.org/socialparticipation/po/>}. |
|  |

|  |
| --- |
| \begin{enumerate}[leftmargin=0cm] |
|  |

|  |
| --- |
| \item Retrieve the number of participants:\\ |
|  |

|  |
| --- |
| \textttt{SELECT (COUNT(DISTINCT ?author) as ?c) WHERE \{ |
|  |

|  |
| --- |
| ?author a po:Participant . \} } |
|  |

|  |
| --- |
| \item Retrieve the number of relations, be them interactions or |
|  |

|  |
| --- |
| friendships:\\ |
|  |

|  |
| --- |
| \textttt{SELECT (COUNT(?interaction) as ?c) WHERE \{\\ |
|  |

|  |
| --- |
| \h \{ ?interaction a po:Friendship \} UNION \{ ?interaction |
|  |

|  |
| --- |
| a po:Interaction \} UNION\\ |
|  |

|  |
| --- |
| \h \{ ?interaction po:retweetOf |
|  |

|  |
| --- |
| ?message \} UNION \{ ?interaction po:replyTo ?message |
|  |

|  |
| --- |
| \}\\ |
|  |

|  |
| --- |
| \h UNION \{ ?interaction po:directedTo ?participant |
|  |

|  |
| --- |
| \}\\ \} } |
|  |

|  |
| --- |
| \item Retrieve all text produced by a specific user:\\ |
|  |

|  |
| --- |
| \textttt{SELECT (CONCAT(?text) as ?texts) WHERE \{\\ |
|  |

|  |
| --- |
| \h ?activity po:author <user\\_uri> . ?activity po:text ?text .\\ |
|  |

|  |
| --- |
| \}} |
|  |

|  |
| --- |
| \item List 1000 users (URIs and names) with the most friendships and the number of |
|  |

|  |
| --- |
| friendships in descending order by the number of friendships:\\ |
|  |

|  |
| --- |
| \textttt{SELECT DISTINCT ?participant (COUNT(?friendship) |
|  |

|  |
| --- |
| as ?c) WHERE \{\\ |
|  |

|  |
| --- |
| \h ?friendship a po:Friendship . ?friendship po:member ?participant . \\ |
|  |

|  |
| --- |
| \} ORDER BY DESC(?c) LIMIT 1000} |
|  |

|  |
| --- |
| \item Retrieve text messages with the word ``pineapple'' (case insensitive):\\ |
|  |

|  |
| --- |
| \textttt{SELECT ?text WHERE \{ \\ |
|  |

|  |
| --- |
| \h ?activity po:text ?text . FILTER regex(?text, 'pineapple', 'i')\\ |
|  |

|  |
| --- |
| \}} |
|  |

|  |
| --- |
| \item List participants and respective full names whose name has the substring ``Amanda'':\\ |
|  |

|  |
| --- |
| \textttt{SELECT DISTINCT ?participant ?name WHERE \{\\ |
|  |

|  |
| --- |
| \h ?participant po:observation ?obs . ?obs po:name ?name .\\ |
|  |

|  |
| --- |
| \h FILTER regex(?name, 'Amanda', 'i') \\ |
|  |

|  |
| --- |
| \}} |
|  |

|  |
| --- |
| \item Return all pairs of friends of a participant which are friends themselves:\\ |
|  |

|  |
| --- |
| \textttt{SELECT DISTINCT ?friend1 ?friend2 WHERE \{\\ |
|  |

|  |
| --- |
| \h ?friendship1 po:member <participant\\_uri> . ?friendship1 po:member ?friend1 .\\ |
|  |

|  |
| --- |
| \h ?friendship2 po:member <participant\\_uri> . ?friendship2 po:member ?friend2 .\\ |
|  |

|  |
| --- |
| \h ?friendship3 po:member ?friend1 . ?friendship3 po:member ?friend2 .\\ |
|  |

|  |
| --- |
| \}} |
|  |

|  |
| --- |
| \item Return all interactions from replies in a snapshot:\\ |
|  |

|  |
| --- |
| \textttt{SELECT ?from ?to WHERE \{\\ |
|  |

|  |
| --- |
| \h ?message1 po:snapshot <snapshot\\_uri> . ?message2 po:replyTo ?message1 .\\ |
|  |

|  |
| --- |
| \h ?message1 po:author ?from . ?message2 po:author ?to .\\ |
|  |

|  |
| --- |
| \}} |
|  |

|  |
| --- |
| \end{enumerate} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{License issues} |
|  |

|  |
| --- |
| % Facebook, Twitter, IRC, Gmane, Participa, CD, AA. |
|  |

|  |
| --- |
| The database presented in this thesis is released under public domain. |
|  |

|  |
| --- |
| Computer scripts are in git repositories and PyPI Python packages, also under public domain. |
|  |

|  |
| --- |
| Although most data is already in open licenses (Twitter, Email, Participabr, Cidade Democrática, and AA data), IRC and Facebook data was collected |
|  |

|  |
| --- |
| and donated by the individuals which yielded the data. |
|  |

|  |
| --- |
| This rises the understanding of the right to study such data as the right to access the self, |
|  |

|  |
| --- |
| in line with anthropological endeavors~\cite{antphy,antphy2}. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| \subsection{Data-driven ontology synthesis} |
|  |

|  |
| --- |
| OWL Ontologies are critical tools to describe taxonomies and the |
|  |

|  |
| --- |
| structure of knowledge. |
|  |

|  |
| --- |
| Most ontologies are created by domain experts even though there often is data they |
|  |

|  |
| --- |
| organize that is given by a software system and which has a predefined |
|  |

|  |
| --- |
| structure. |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
| We developed a simple ontology synthesis method that probes |
|  |

|  |
| --- |
| the ontological structure in data with |
|  |

|  |
| --- |
| SPARQL queries and post-processing. |
|  |

|  |
| --- |
| The results are OWL code and diagrams which are |
|  |

|  |
| --- |
| exemplified in the Appendix~\ref{losd} and |
|  |

|  |
| --- |
| available in an article~\cite{losd}. |
|  |

|  |
| --- |
| The method can be extended to comprise further OWL axioms and restrictions, |
|  |

|  |
| --- |
| but is currently performed to fit present needs with maximum simplicity. |
|  |

|  |
| --- |
| The present needs are limited to informative figures and |
|  |

|  |
| --- |
| the steps implemented are as follows: |
|  |

|  |
| --- |
| \begin{enumerate}[leftmargin=0cm] |
|  |

|  |
| --- |
| \item Obtain all distinct classes with the query:\\ |
|  |

|  |
| --- |
| \textttt{SELECT DISTINCT ?class\\_uri WHERE \{ ?s a ?class\\_uri \}} |
|  |

|  |
| --- |
| \item For each class, obtain the properties that occur as predicates in triples where the subject is an instance of the class:\\ |
|  |

|  |
| --- |
| \textttt{SELECT DISTINCT ?property\\_uri WHERE \{ ?s a <class\\_uri> . ?s ?property\\_uri ?o . \}}\\ |
|  |

|  |
| --- |
| Such properties are used to assert existential and universal restrictions for the class. |
|  |

|  |
| --- |
| \item Compare the total number of individuals (\textttt{?cs1}) of the class (\textttt{class\\_uri}) with |
|  |

|  |
| --- |
| the number of such individuals (\textttt{?cs2}) that are subjects of at least one triple where |
|  |

|  |
| --- |
| the predicate is the property (\textttt{property\\_uri}). |
|  |

|  |
| --- |
| If the numbers match, there is an existential restriction for the class. The queries are:\\ |
|  |

|  |
| --- |
| \textttt{SELECT (COUNT(DISTINCT ?s) as ?cs1) WHERE \{ ?s a <class\\_uri> \}}\\ |
|  |

|  |
| --- |
| \textttt{SELECT (COUNT(DISTINCT ?s) as ?cs) WHERE \{\\ |
|  |

|  |
| --- |
| \h ?s a <class\\_uri>. ?s <property\\_uri> ?o .\\ \}} |
|  |

|  |
| --- |
| \item Find the number of instances which are subjects of triples where the predicate is the property but are not instances of the class. |
|  |

|  |
| --- |
| If there is zero of such instances, there is a universal restriction:\\ |
|  |

|  |
| --- |
| \textttt{SELECT (COUNT(DISTINCT ?s)=0 as ?cs) WHERE \{\\ |
|  |

|  |
| --- |
| \h ?s <property\\_uri> ?o . ?s a ?ca . FILTER(str(?ca) != 'class\\_uri')\\ \}} |
|  |

|  |
| --- |
| \item To keep a record of the restrictions (and occurring triples), get all object classes or datatypes where the subject is an instance of the class and the predicate is the property:\\ |
|  |

|  |
| --- |
| \textttt{SELECT DISTINCT ?co (datatype(?o) as ?do) WHERE \{\\ |
|  |

|  |
| --- |
| \h ?s a <class\\_uri>. ?s <property\\_uri> ?o . OPTIONAL \{ ?o a ?co . \}\\ |
|  |

|  |
| --- |
| \}} |
|  |

|  |
| --- |
| \item Obtain all distinct properties:\\ |
|  |

|  |
| --- |
| \textttt{SELECT DISTINCT ?p WHERE \{ ?s ?p ?o \}} |
|  |

|  |
| --- |
| \item Check if each property is functional, i.e. if it |
|  |

|  |
| --- |
| occurs at most once with each subject. |
|  |

|  |
| --- |
| This is performed by counting the objects and further verifying |
|  |

|  |
| --- |
| that they are at most one. The query is:\\ |
|  |

|  |
| --- |
| \textttt{SELECT DISTINCT (COUNT(?o) as ?co) WHERE \{ ?s |
|  |

|  |
| --- |
| <property\\_uri> ?o \} GROUP BY ?s} |
|  |

|  |
| --- |
| \item For each property, find the incident range and domain with the |
|  |

|  |
| --- |
| queries:\\ |
|  |

|  |
| --- |
| \textttt{SELECT DISTINCT ?co (datatype(?o) as ?do) WHERE \{\\ |
|  |

|  |
| --- |
| \h ?s <property\\_uri> ?o . OPTIONAL \{ ?o a ?co . \}\\\}} \\ |
|  |

|  |
| --- |
| and \\ |
|  |

|  |
| --- |
| \textttt{SELECT DISTINCT ?cs WHERE \{ ?s <property\\_uri> ?o . ?s a ?cs . \}} |
|  |

|  |
| --- |
| \item Render diagrams as explained in the next section and in the Supporting Information file. |
|  |

|  |
| --- |
| \end{enumerate} |
|  |

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |

* [Contact GitHub](https://github.com/contact)
* [API](https://developer.github.com)
* [Training](https://training.github.com)
* [Shop](https://shop.github.com)
* [Blog](https://github.com/blog)
* [About](https://github.com/about)
* © 2017 GitHub, Inc.
* [Terms](https://github.com/site/terms)
* [Privacy](https://github.com/site/privacy)
* [Security](https://github.com/security)
* [Status](https://status.github.com/)
* [Help](https://help.github.com)